Nyelvészeti problémák a szabadalmak feldolgozásában

Vincze Veronika[1], Nagy Ágoston[1], Klausz Ágnes[1], Almási Attila[1], Kiss Márton[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{vinczev, nagyagoston, aklausz, mkiss}@inf.u-szeged.hu,
vizipal@gmail.com

Kivonat: A szabadalmak számos olyan sajátossággal bírnak, amelyek azok nyelvi elemzését – az általános tématerületű szövegekhez képest – jelentősen megnehezítik. Szintaktikailag bonyolult felépítésű szerkezetek, beágyazott mondatok, összetételek és felsorolások szép számmal találhatók bennük, igen sok bennük a visszautalás (anafora), és az elliptikus tagmondatok, vonatkozó mellékmondatok és utómódosítók használata is jellemző. A szabadalmak szókincse is jellegzetes: a terminus technicusokon kívül bizonyos szófordulatok jelenléte is tipikusnak mondható. Mindezen jellemzőkből adódó problémák kezelésére különféle szabályalapú módszereket dolgoztunk ki, melyeket az előadásban ismertetünk.

1 Bevezetés

Az ALL és a Szegedi Tudományegyetem egy közös projekt keretében vállalta egy szemantikus keresőrendszer kifejlesztését, amely elsődlegesen az angol és magyar nyelvű szabadalmakban való keresést célozza meg, ugyanakkor a készülő rendszer könnyen adaptálható lesz más területekre is. Mivel a szabadalmak rendkívül sok tudományterületet fednek le, melyek mindegyike sajátos jellemzőkkel bír (mind stilisztikai, mind terminológiai szempontból, mind pedig a szabadalmak felépítését tekintve), a projekt keretein belül egy adott osztályozási jelzettel ellátott szabadalmak feldolgozására összpontosítunk, nevezetesen az A61K (orvostudományi) osztályra.

A szabadalmak számos olyan sajátossággal bírnak, amelyek azok nyelvi elemzését – az általános tématerületű szövegekhez képest – jelentősen megnehezítik. Az előadásban e sajátosságokat, az ezekből adódó problémákat és a rájuk adott megoldásokat ismertetjük.

2 A szabadalmak felépítése

A szabadalmak egységes szerkezettel bírnak. A címlap tartalmazza az úgynevezett bibliográfiai adatokat, amelyben megtalálható többek között a szabadalom iktatási száma, a benyújtás időpontja, a szerzők és a feltalálók neve. Az első oldalon szerepel még a találmány néhány soros összefoglalója, amelyet ábrákkal is ki lehet egészíteni.

Itt található a cím is, amely meghatározza a találmány tárgyát, majd a leíró részben annak pontos jellemzőit fejtik ki a szerzők különös tekintettel a találmánnyal megoldandó feladatra, az alkalmazási területekre, példákkal, ábrákkal, táblázatokkal szemléltetve. Az igénypontok pedig a szabadalmak oltalmi körét határozzák meg, azaz azt, hogy mit szeretnének a feltalálók levédetni.

A találmányt az úgynevezett főigénypont azonosítja a legáltalánosabban. A főigénypontban megtalálható a találmánynak a célul kitűzött feladat megoldásához elengedhetetlenül szükséges minden jellemzője (l. [7]). Emiatt a továbbiakban elsődlegesen a főigénypontok nyelvi feldolgozására összpontosítunk.

A főigénypont szerkezete eléggé kötött. Ez már abból is adódik, hogy a főigénypont hossza csak egy mondat lehet: a legtöbb problémának ez a forrása, mert mindent ebbe az egy mondatba próbálnak beletömöríteni. A főigénypont mindig azzal kezdődik, hogy milyen kategóriába tartózik a levédetni kívánt szabadalom, például módszer, eljárás, eszköz, összetétel. Eztán következik ezek kifejtése: milyen lépésből/anyagokból áll a főigénypont elején említett dolog, és ezeket az alpontokat rekurzívan továbbfejtik.

3 A szabadalmak nyelvi jellemzői

Mint már említettük, a szabadalmak terminológiai és stilisztikai szempontból is eltérnek az általános doménből vett szövegektől. Mind a magyar, mind az angol szabadalmakra jellemző, hogy nyelvezetük tömör, lényegre törő. Szintaktikailag bonyolult felépítésű szerkezetek, beágyazott mondatok, összetételek és felsorolások szép számmal találhatók bennük. A megfogalmazásban pontosságra törekednek a szerzők, igyekeznek kimerítő leírást adni a találmányról, ugyanakkor megfigyelhető az a tendencia is, hogy – az esetleges későbbi jogviták elkerülése végett – bizonyos általánosító stratégiákat alkalmaznak, így lehetővé válik a jellemzők és az alkalmazási területek bővítése, illetve a későbbiekben esetleg relevánssá váló esetek hozzáadása („beleértése" a szabadalomba) [7]. Ilyen nyelvi stratégiára hozunk néhány példát:

- a kimerítőnek látszó felsorolások végén szereplő *stb*.;
- a felsorolások előtt szereplő *pl.* vagy *például*;
- megengedő *vagy* használata;
- általános jelentéstartalmú határozók használata (*rendszerint*, *általában*).

E stratégiák némileg párhuzamot mutatnak a bizonytalanságot jelölő kifejezésekkel (angol terminológiával élve a hedge, illetve weasel kifejezésekkel [2]), míg azonban például a Wikipédia szócikkein belül ezen általánosító, kétértelmű és nem kimerítő leírást adó kifejezések használata nemkívánatosnak minősül, addig a szabadalmak nyelvezetében a fenti okok miatt ez teljességgel megszokott stratégia.

Mivel a főigénypontnak tartalmaznia kell minden szükséges, a szabadalom lényegét érintő információt, továbbá a hagyományoknak megfelelően a főigénypont egyetlen mondatból áll, ezért nem várható el, hogy a főigénypontot egy egyszerű, könnyen feldolgozható mondat alkossa [7]. Szintaktikai szempontból jellemezve a mondatokat elmondhatjuk, hogy igen hosszú, többszörösen összetett mondatok alkotják a szaba-

dalmak szövegét – egy-egy főigénypont (azaz egy mondat) akár több oldal hosszúságú is lehet. Ebből adódóan igen sok bennük a visszautalás (anafora), és az elliptikus tagmondatok, felsorolások, vonatkozó mellékmondatok és utómódosítók használata is jellemző. A mondatok pontos szintaktikai elemzését a fentiek mellett az is nehezíti, hogy a központozás nem túl következetes. A fentiek miatt [7] szerint a szabadalmak külön nyelvtannal (szintaxissal) bírnak, mely nem esik egybe a(z angol) nyelvtannal.

A szabadalmak szókincse is jellegzetes: a terminus technicusokon kívül bizonyos szófordulatok (*azzal jellemezve*) jelenléte is tipikusnak mondható, melyek nem feltétlenül találhatók meg egy általános célú szótárban, így ezeket külön fel kell venni, illetve a kezelésükre külön szabályokat kell írni. A szabadalmak értelmezését az is megnehezítheti, hogy – mivel a leírt találmány új – a találmány leírására használt szavak is új értelmezésben használtatnak a szabadalomban [7].

4 Nyelvi problémák a szabadalmakban

A szabadalmak nyelvi sajátosságaiból adódó, az általános doménre felkészített nyelvi elemzők számára [5] problémát jelentő esetek a következők:

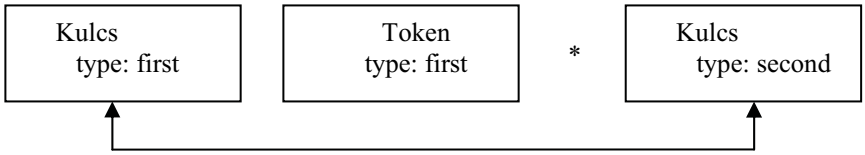
- rendkívül hosszú mondatok (kulcsok és utómódosítók)
- adjunktumok
- sajátos fordulatok
- összetételek
- felsorolások
- kvantitatív szerkezetek
- kémiai névelemek

A fenti problémák kezelésére különféle szabályalapú módszereket dolgoztunk ki, melyeket az alábbiakban ismertetünk részletesen.

4.1 Kulcsok

Egy szabadalom főigénypontja általában egy többszörösen összetett, nagyon nehezen elemezhető mondat sok alá- és mellérendeléssel. Ezeknek a nem ritkán több mint száz szavas mondatoknak a gépi elemzése a jelenlegi elemzők segítségével nem lehetséges. Olyan megoldást kellett találnunk, amely segítségével e mondatokat olyan elemi mondattöredékekre tudjuk bontani, melyek elemezhetőek gépi algoritmusok segítségével. Ezért az utómódosítók, valamint a mellékmondatok kezdetét **kulcs**okkal jeleztük.

Kulcs alatt általánosan a feldolgozott szövegnek azokat a szakaszait értjük, ahol a módosító-módosított főnév viszony jelenléte *pusztán formai alapon* felismerhető. A kulcsok egy első és egy második részből épülnek fel.



1. ábra. A kulcsok felépítése.

- **Egyszerű kulcs:** Az egymást követő kulcsok jelölésére szolgál abban az esetben, ha a kulcs első részéhez nem kapcsolódik távoli második típusú kulcs. Például: *substance which*, *group consisting*.
- **Összetett kulcs:** Összetett kulcsról beszélünk, ha a kulcs első és második tagja nem közvetlenül követi egymást, vagy a kulcs első részéhez több második rész is tartozik. Például: *the **process comprising** the steps of deforming the films (18) to form a multiplicity of recesses (16), **filling** the recesses*.
- **Beágyazott kulcs:** Minden olyan esetben alkalmazandó, ahol nem érvényesíthető a következő szabály: „Összetett kulcs második részét mindig az előtte álló összetett kulcs első részéhez kell kötni”. A beágyazott kulcsok egymással sorfolytonosan balról jobbra, kettesével kötendők össze és feldolgozásuk megelőzi az összetett kulcsét. Például: *A **method** for the treatment of systemic infection **diseases**, such as pneumonia, tuberculosis, peritonitis, endocarditis, pyelonephritis, meningitis or septicemia, **caused** by bacterial or protozoal infection, **comprising**:*.

A kulcsokat két osztályba soroljuk felismerhetőségük alapján:

1. Egymást követő kulcs, ezen kulcspárok egésze (első és második részük is) egyből felismerhető. A következő esetekben tekinthető kulcsnak két egymást követő token (a lenti felsorolásban a Stanford szófaji egyértelműsítő [5] jelölésrendszerét használjuk):

   - N + postModifier
   - N + to + VB/VBP
   - N + JJ + Prep
   - N + (WDT|WP|WP$)

2. Csak az elemzés későbbi részében felismerhető kulcspárok, ezen kulcspároknál csak a kulcs második része ismerhető föl pusztán formai jelek alapján. E kulcsok első része az elemzés későbbi részében ismerhető föl, illetve keresendő meg. A következő esetekben tekinthető kulcsnak (kulcs második részének) egy token:

   - **whose**

- **which,** ha előtte **,** vagy **;** van, vagy **and** tokenek állnak
- Minden VBN szófaji kóddal rendelkező token, ha megelőzi egy **,** vagy **;**
- A következő szavak: **comprising|having|consisting|being|including ,** ha megelőzi őket egy **,** vagy **;** vagy az **and**

4.2 Adjunktumok

A köznyelvhez képest szerencsére igen kevés az adjunktumok száma a szabadalmak igen kötött nyelvezetének köszönhetően (csak azt mondják, ami feltétlenül szükséges, azt viszont pontosan). Néhány esetben azonban különös figyelmet igényelt az adjunktumok kezelése.

Az *optionally* gyakorlatilag *vagy*-szerű logikai operátorként viselkedik (valami vagy megtörténik, vagy nem), ezért a szemantikai elemzés során erre hangsúlyt kell fektetni. Egy példa:

*C.sub.6-C.sub.10-arylthio which is **optionally** substituted by nitro, amino, C.sub.1-C.sub.6-alkyl or C.sub.1-C.sub.4-alkoxy*

A példában a *C.sub.6-C.sub.10-arylthio* helyett állhat vagy *nitro*, vagy *amino*, vagy *C.sub.1-C.sub.6-alkyl* vagy *C.sub.1-C.sub.4-alkoxy*.

Egy másik lehetséges problémaforrás, hogy a szabad határozó néha az ige és a vonzata között helyezkedik el:

*consisting **essentially** of a purified mineral composition and optional excipients*

Ez a vonzatkeret illesztése miatt okozhat problémákat, de néhány szabály segítségével áthidalható, szemantikai szinten pedig az ilyen módon az igéhez kapcsolódó legtöbb határozó jelentése elhanyagolható a mondat szempontjából.

A PP-bővítmények (*during a sport activity, without a tabletting excipient*…) vagy az előtte levő NP részei (ill. a főnévi fej bővítményei), vagy pedig az igéhez kapcsolódnak. Ennek eldöntése igen nehéz, sokszor még az ember számára sem egyértelmű. A főnevekhez készítendő vonzatkerettárat kellett ilyen esetekben segítségül hívni (ha a főnévi fejhez egy adott prepozíciót tartalmazó PP kapcsolódik, akkor a főnév bővítményeként kezeljük, ha nem, akkor az igéhez tartozóként), vö. [4].

Bizonyos, jelzőket módosító határozószavak (*pharmaceutically, substantially, dermatologically, therapeutically*…) gyakran kollokációszerűen viselkednek:

*a **dermatologically acceptable** carrier*
*a **therapeutically effective** amount of a compound of Formula I*
*a **pharmaceutically acceptable** salt thereof*

Ezeket egységként vettük fel a szótárban.

4.3 Sajátos fordulatok

A szabadalmak szókincsének jellegzetes elemei bizonyos szófordulatok (*said*, *a plurality of*, *azzal jellemezve…*), melyek nem feltétlenül találhatók meg egy általános célú szótárban, így ezeket külön fel kell venni, illetve a kezelésükre külön szabályokat kell írni. Például a fenti *said* jelző anaforikusan utal vissza egy, a szabadalmi igénypont szövegében már korábban megemlített entitásra, így anaforikus elemként érdemes kezelni.

Az *a plurality of* típusú szerkezetek szemantikailag átlátszóak, noha szintaktikailag a *plurality* számít a kifejezés fejének, szemantikai szinten az *of* prepozíció bővítménye játszik csak fontos szerepet:

A vitamin supplement to temporarily enhance the abilities of a individual during a sport activity comprising a plurality of B family vitamins and one or more other vitamins, minerals, and/or natural ingredients.

Ebből következően a mondat szemantikai reprezentációjában az *a plurality of* nyelvi kifejezés nem is szerepel.

Az *azzal jellemezve* típusú szófordulatokat külön elemként szerepeltetjük a szótárban.

4.4 Összetételek

Az elemzés során problémát okozhatnak a halmozott NP-szerkezetek, ezen belül is különösen az előmódosítók. Mint fentebb említettük, a szabadalmi szövegekre kifejezetten jellemző a tömörség, az informativitásra való törekvés, ami – többek között – a rendkívül hosszú mondatokban, szószerkezetekben nyilvánulhat meg. Ráadásul az angol nyelvben a főnévi előmódosítók számának csupán az érthetőségi korlátok szabnak határt. A több, közvetlenül egymás után álló főnév a gépi elemzés során elsősorban szegmentálási problémát jelenthet.

Többek között az N + ADJ + N szerkezetű magNP-k okozhatnak ilyen problémát, mivel a szerkezeti elemzésük többféleképpen történhet. Alapvetően kétféle variáció állhat fenn: a középső elem, azaz a melléknévi alak vagy az előtte álló főnévhez kapcsolódhat szorosabban, vagy az utána állóhoz. Az utóbbi esetben az N + ADJ szerkezetű NP-nek az első főnév az előmódosítója: [N + [ADJ + N]. A gépi elemző általában ezt a szegmentálási variációt használja alapértelmezésként.

Azonban vannak esetek, amikor az N + ADJ + N szerkezet mellékneve – bár szintén az *utána* álló főnév előmódosítója – az *előtte* álló főnévhez szorosabban kapcsolódik, mivel a vele alkotott jelzői módosító feje. (Itt az első főnév az előmódosító előmódosítója):

[[N + ADJ] + N], pl. [[*silicone conditioning*] *oil*].

Ilyen esetekben a szintaktikai elemzés során a melléknév *után* kell részekre bontani az NP-t. (Amennyiben névelő áll a második főnév előtt, egyértelmű, hogy a melléknevet az *előtte* álló főnévhez kell kapcsolni.)

A szóban forgó melléknévi alakok lehetnek *–ing* végződésűek, illetve *past participle* alakúak. Az előbbiek többnyire tárgyas igéből képzett folyamatos melléknévi igenevek, pl. *containing*, vagy tárgyas igéből képzett melléknevek, pl. *(pH-)responsive, (bio-)absorbable,* de lehetnek egyszerű melléknevek is, pl. *(sodium-)free.* A *past participle* alakúak szintén tárgyas igéből képzettek: *(diabetes-)associated, (lipoprotein receptor-)related.*

A fentebbieken kívül kétértelműek lehetnek még az ADJ + ADJ + N szerkezetű szóösszetételek is, amelyeket [ADJ + [ADJ + N]] szerkezetként (pl. *substituted lower alkyl, inorganic metal oxide*) és [[ADJ + ADJ] + N] szerkezetként (*vascular-related diseases*) is lehet értelmezni.

A melléknevet tartalmazó előmódosítókban az első elem lehet számosságra utaló elem is, ami szintén azt a problémát veti fel, hogy hova kapcsoljuk az utána álló melléknevet abban az esetben, ha nincs kötőjel az elemek között, pl. *penta-substituted C1-C12 alkyl, three- to seven-membered alkylene bridge.*

4.5 Felsorolások

Mivel a szabadalmak főigénypontjai egymondatosak lehetnek csak, ezért a szerzők abba az egy mondatba próbálnak mindent belesűríteni. Ez a felsorolások kezelésének tekintetében is sok bonyodalmat okoz. A felsorolásokat formailag viszonylag könnyű felismerni, mert elemeit vessző, pontosvessző vagy kötőszó választja el (habár sok esetben ez hiányzik). A szintaktikai elemzés szempontjából viszont gyakran nehéz eldönteni, hogy a felsorolást elválasztó elemek után található szó vagy szócsoport minek a bővítménye. Ez amiatt történhet meg, hogy a főösszetevők felsorolása mellett párhuzamosan történik meg az azokban található alösszetevők leírása, amelyek szintén tovább bonthatók. Esetenként így akár 3-4 szint mélységű is lehet egy-egy felsorolás. Általában a vesszővel azonos szinten lévő elemeket sorolunk fel, a pontosvesszőt pedig legalább egy szinttel megy feljebb – de a ”legalább egy” és az ”azonos szinten” sajnos nem elég pontos támpont egy parser létrehozása szempontjából, mert kivételek is lehetnek. Erre példa az alábbi szabadalomrészlet:

R1 and R2 are each selected independently from the group consisting of hydrogen, hydroxyl, amino, ..., alkoxy of 1-6 carbon atoms, alkylthio, aryloxy, ...

A fenti példában az tapasztalható, hogy a *consisting* vonzata a *hydrogen, hydroxyl, amino, alkoxy of 1-6 carbon atoms, aryloxy* stb. Ez számunkra teljesen evidens, de a felsorolásokkal kapcsolatban felállított szabályok szerint a parser logikusan az *alkylthio* és az azt követő felsoroláselemeket az *alkoxy* szóhoz köti, pedig valószínűleg azok is a *consisting* szóhoz tartoznak. Az *atoms* utáni vessző tehát nem azonos szintet, hanem egy szinttel feljebb való ugrást feltételez. A problémán itt még az sem segítene, ha minden, felsorolásban található elem előtt megismételjük a prepozíciót, mert itt mindkét esetben az *of* lenne az.

A felsorolások végén található *and* vagy *or* kötőszó pedig azt jelenti, hogy az adott felsorolás utolsó eleme fogja követni. Ez sok esetben igaz, de találtunk egy többszörösen mellérendelt mondatkezdetet is:

A means for allaying drunkenness, preventing and removing alcohol intoxication and hangover syndrome and ***a method*** *for allaying drunkenness, preventing and removing alcohol intoxication and hangover syndrome by using this means, comprising:*

A fenti példában a *removing* utáni felsorolás okoz problémát: a *preventing* és *removing* tárgyas vonzata az *alcohol intoxication* és a *hangover syndrome*. Azonban ezekhez még hozzá van kötve szintén az *and* kötőszóval a *method* is, amely az elemző számára természetesen ugyanolyan, mint az *alcohol intoxication*, így azokhoz köti testvérként. Itt semmi sem jelzi a feljebb ugrást, ami ráadásul kétszintű: nem a *means for* vonzata a *method*, hanem a gyökérhez köthető a *means* mellé.

4.6 Kvantitatív szerkezetek

A biokémiai szabadalmakban fontos szerepük van a mennyiségjelzőknek, amelyek feladata, hogy a főigénypontokban minél pontosabban leírják egy kémiai összetétel összetevőinek pontos mennyiségét. Mivel a főigénypontok a mérvadóak a szabadalmaztatás során, a szerzők nemcsak az előbb említett pontosságra törekednek, hanem arra is, hogy hasonló összetételt se lehessen alkalmazni, így gyakran használnak olyan szerkezeteket, amelyek az összetevők mennyiségét a *körülbelül* előtaggal módosítják. Így a főigénypontokban egyszerre jelenik meg a pontosság igénye, és a mennyiségmegjelölések kis mértékű elhomályosítása (vö. 3. fejezet).

A szabadalmak mennyiségei rögzített szerkezettel rendelkeznek: általában *-tól/-ig* tartományt fejeznek ki, például *from about 1 gram to about 5 grams of Arginine*. Az ilyen típusú mennyiségjelzők szintaktikai szempontból nem okoznak problémát: általában mindegyik egy megadott mintára illeszkedik, így azok kinyerése viszonylag könnyen megoldható. Szemantikai szempontból viszont az ilyen típusú szerkezetek problémát okozhatnak. Ha egy szabadalmi keresőbe beírjuk, hogy olyan összetételeket keresünk, amelyben *0,5 gramm Arginine* található, akkor az beleesik-e a fent említett példába, azaz *a kb. 1 grammtól kb. 5 grammig terjedő* tartományba? A *körülbelül* szónak így meg kell adni egy viszonylag széles tartományt, amelybe biztosan belefér a keresett elem, de felesleges találatokat nem ad. Ennek a problémának a megoldása további fejlesztések eredményeképpen várható.

A mennyiségjelzős szerkezetek esetében a felismerési problémát az okozza legtöbbször, hogy a mennyiséget kifejező tag túl messzire kerül a hozzá tartozó főnévtől, így azok összekötése nehézzé válik. Vannak olyan esetek, amikor csak a *be* ige ragozott alakjai kerülnek be a mennyiségjelző és a hozzá tartozó főnév közé:

the weight ratio of xanthan to guar gum [being] from 1:3 to 1:10
the weight ratio of crystals to carrier [is] 2-99%

Ezen esetekben a *be* elhagyásával a mennyiségjelző könnyen összeköthető. Azonban vannak olyan esetek, ahol a mennyiségjelzők és a hozzájuk tartozó főnevek nagyon messzire elkerülnek egymástól. Az alábbi két példa is ezt szemlélteti:

the sodium bicarbonate being incorporated in the toothpaste in an amount of at least 60% by weight

the ratio of the components is as follows (wt. %): TBL natural minerals 33-62 vegetable stock 34-61 water the balance.

Az első esetben a *legalább 60 tömeg%* a nátrium-bikarbonátra vonatkozik, de közéjük beékelődik még az, hogy ez az arány miben található, nevezetesen a fogkrémben. A második egy elég extrém példa, és szerencsére ritka is. Itt a mértékegység zárójelben kikerül előre, és egy felsorolásban következik utána az összetevők listája, majd azok mennyisége (már mértékegység nélkül). A természetes ásványok tömegszázaléka 33-62, a zöldségéé 34-61, a többi pedig víz. A felsorolásoknál tovább nehezíti a dolgot, hogy ebben az esetben sincs vessző a felsorolások tagjai között.

Gyakori probléma még, hogy a szöveges formátum nem mindig megfelelő: például táblázatokból egyszerű szövegek keletkeznek, a sorok és oszlopok összemosódásával. Ezekben az esetekben a mennyiségeket még nehezebb összekapcsolni a főnévvel. Erre példa az alábbi táblázat, amelynek szöveges változatát alatta közöljük:

particle size	percentage
5 μm or more and less than 100 μm	5 to 30%
100 μm or more and less than 300 μm	10 to 40%
300 μm or more and less than 500 μm	10 to 50%
500 μm or more and less than 1000 μm	balance

particle size percentage 5 μm or more and less than 100 μm 5 to 30% 100 μm or more and less than 300 μm 10 to 40% 300 μm or more and less than 500 μm 10 to 50% 500 μm or more and less than 1000 μm balance

Ebben a példában a részecskeméréthez tartoznak az alatta lévő elemek, és a százalékhoz az abban az oszlopban található mértékek, a folyó szövegben viszont ezt nehéz összepárosítani.

A kvantitatív szerkezetek felismerésében egy másik nagyobb problémát a létező mértékegységek nagy száma jelenti. További probléma, hogy a mértékegységek gyak-ran rövidített alakjukban szerepelnek, melyek igen gyakran csak 1-2 karakterből áll-nak, ami többértelműségekhez vezethet (pl. az mg betűsor – kis- és nagybetűket nem megkülönböztetve – lehet a magnézium vegyjele is és milligramm is, a C pedig lehet Celsius-fok és a szén vegyjele is, vö. [1, 6]).

4.7 A névelemek annotációja során felmerült problémák

A szabadalmak annotálásakor olyan névelemeket kerestünk, amelyek a kémia területéhez tartoznak, és amelyekre a felhasználó nagy valószínűséggel rákereshet. Három kategóriát vettünk fel: 1) kémiai elemek (nitrogén, oxigén), elemcsoportok (halogének, alkáli földfémek), vegyületek (Na_2O, CaO) és egyéb olyan kifejezések, amelyek

az annotáló számára elég specifikusak voltak ahhoz, hogy ebbe a halmazba kerülhessenek; 2) egyéb, biokémiai szempontból fontos kifejezések: pl. általános anyagnevek (ginzeng, cukor, só stb.), vegyületfajták (szénhidrogének) és egyéb olyan kifejezések, amelyek kémiai szempontból keresőkifejezések lehetnek; 3) konkrét betegségek (Alzheimer-kór, tuberkolózis), betegségcsoportok (gyulladásos betegségek, immunhiányos betegségek) és tünetek (másnaposság).

A kifejlesztett NER modul futásának eredménye a következőkre irányította a figyelmet:

1. A program bizonyos esetekben nem különíti el a névelemek főnévi és jelzői használatát, amire példa az *antibiotic* szó, mely az angolban főnévként és melléknévként is szerepelhet, és a szabadalmakban is kétféleképpen fordul elő (vö. *an* ***antibiotic*** *medication – a total amount of* ***antibiotic*** *and antihistamine*). Az annotálás során a főnévi szerepben lévő elemeket jelöltük.

2. Az annotálás első körében úgy jártunk el, hogy csak azokat az elemeket vettük fel NE-nek, amelyek valamely képlettel (egyértelműen) azonosíthatók voltak. Így fordult elő pl. az anyagnevek esetében, hogy egy adott alakban előforduló szót egyszer NE-nek jelöltünk, más esetben viszont nem. Erre a legjobb példa az *alcohol* szó, mely egyes szabadalmakban valamilyen kémiai szempontból jól beazonosítható vegyület részét képezi (*cetylstearyl alcohol*), máskor viszont csupán mint szeszesital szerepel (pl. az *alcohol intoxication*ben).

A szabadalmakban való keresés és az annotálási elvek nagyobb fokú összehangolása érdekében a jelölési elveket módosítottuk, két kémiainévelem-kategóriát vettünk fel (lásd fentebb), s így az *alcohol*t már minden esetben jelöltük.

3. Többször előfordult, hogy a program – pl. a szabadalmakban előforduló helyesírási hibák miatt – nem megfelelően szegmentált bizonyos elemeket (pl. *…alkarylamino, fluoro, chloro,* ***bromo iodo*** *and trifluoromethyl…*), ezért két, egyébként különálló NE-t egynek tekintett. Ezekben az esetekben a jelölést a valós tartalomból kiindulva (és a nyelvhelyességnek megfelelően) végeztük el.

4. Szófaji problémák:

a) A program minden olyan elemet, amely a szótárjában NE-ként szerepel, alkalmas jelöltnek tekint és kiemel. Pl. a ***water****-soluble*, ***sodium****-free*, ***wax****-like* (vízoldékony, nátriummentes, viasszerű) kifejezések a magyarban egyértelműen nem számítanak névelemnek, második tagjuk pedig az úgynevezett HALFLEX melléknevek közé tartozik [8]. A program úgy jár el, hogy ha talál NE-t, és az kötőjellel kapcsolódik egy másik elemhez, akkor az NE határát kiterjeszti, és annak részeként kezeli a kapcsolódó elemet is, ami ezekben az esetekben nem megfelelő eljárás. A kézi annotálás során ezeket az elemeket nem jelöltük.

b) Egy másik esete annak, hogy a program NE-ként jelöl meg bizonyos, egyébként nem jelölendő elemeket pl. a *carboxylic* és az *enantiomeric* jelzők, amelyekben szerepel egy-egy, a szótárprogramba felvett NE, a *carboxyl* vagy az *enantiomer*, de ami-

att, hogy a program kiterjesztéses elven működik, a teljes kifejezést NE-nek jelöli. Az annotálás során ezeket az elemeket nem jelöltük.

c) Harmadik példa a nem megfelelő jelölésre az *O-glycosidically*. A szótárprogram a nagy *O*-t NE-ként kezeli, és mivel az a) ponthoz hasonlóan, kötőjellel kapcsolódik az utána következő taghoz, a kettőt egy NE-nek veszi, ami szintén nem megfelelő, mivel a teljes kifejezés egy határozószó. A kifejezés itt sem lett megjelölve.

5 A korpusz

A nyelvészeti problémák feltárásához és a kidolgozott algoritmusok és módszerek ellenőrzéséhez nélkülözhetetlen volt összeállítanunk és kézzel annotálnunk egy korpuszt. A korpusz 313 szabadalmat tartalmaz az IPC osztályozási rendszer A61K besorolású szabadalmai közül. Mivel a kutatás jelen fázisában a szabadalmak fő igénypontjait tanulmányozzuk így ezekben jelöltük be kézzel az alábbiakat: 1) kvantitatív szerkezetek mintái; 2) perdurant jelentésű kifejezések; 3) kulcsok; 4) kémiai névelemek és 5) felsorolások és felsorolásjelzők.

A korpuszon az annotálás Microsoft Wordben történt, majd e dokumentumokat konvertáltuk TXT-be és az annotációkat pedig UIMA-ba [3]. Így könnyen elemezhettük és felhasználhattuk a kézzel jelölt korpuszt.

6 Eredmények

A kulcsok felismerésére létrehozott program működésének kidolgozásához, valamint a program ellenőrzésére egy 60 szabadalomból álló korpuszban jelöltük be kézzel a kulcsokat. A mintakorpusszal összehasonlítva a kulcsok azonosítására kidolgozott eljárást az alábbi mérőszámokat kaptuk.

1. táblázat: A kulcsok felismerésének eredményei.

	Pontosság	**Fedés**	**F-mérték**
Kulcsok megszorítás nélkül (teljes kulcs):	75.47%	75.59%	75.53%
Csak a kulcs első része:	70.61%	71.09%	70.85%
Csak a kulcs második része:	78.27 %	78.042 %	78.16%

A fenti értékekből is látszik, hogy az algoritmus a kulcsok első felének detektálásakor hibázik többet, míg a kulcsok második felét valamivel jobban képes detektálni. A kapott értékek növelése egy bizonyos szintig megoldható további szabályok bevezetésével. További eredményeink: a kémiai névelemek felismerésében 95,25%-os F-mértéket, míg a magNP-k azonosításában 92,59%-os F-mértéket értünk el.

7 Összegzés

A tanulmányban bemutattuk a szabadalmak nyelvi sajátosságait és az azokból fakadó elemzési problémákat. Utóbbiakra számos szabályalapú megoldást dolgoztunk ki, melyek segítségével az elemző algoritmusunk mind pontosság, mind fedés terén (azaz F-mértéket tekintve is) számottevő javulást mutatott. A jövőben az algoritmus további tökéletesítése, illetve a most még nem megoldott problémák (pl. felsorolások) kielégítő kezelése a célunk.

Köszönetnyilvánítás

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Agatonovic, M., Aswani, N., Bontcheva, K., Cunningham, H., Heitz, T., Li, Y., Roberts, I., Tablan, V.: Large-scale, Parallel Automatic Patent Annotation. In: Proceedings of 1st International CIKM Workshop on Patent Information Retrieval - PaIR'08. Napa Valley, California, USA (2008)
2. Farkas, R., Vincze, V., Móra, Gy., Csirik, J., Szarvas, Gy.: The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, Uppsala (2010) 1–12
3. Kiss M., Nagy Á.: Egy nyelvészeti UIMA folyamat a kézi annotálástól az eredmények megjelenítéséig. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 362–364
4. Klausz Á., Vincze V., Nagy Á., Almási A.: Vonzatkeretek vizsgálata orvostudományi tárgyú, angol nyelvű szabadalmi szövegeken. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 180–189
5. Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (2003) 423–430
6. Nyilas S., Németh G., Almási A.: Szótáralapú kémiai NE-felismerő rendszer. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 379–383
7. Osenga, K.: Linguistics and patent claim construction. Rutgers Law Journal Vol. 38, No. 61 (2006) 61–108
8. Vincze V., Lucza M., Csendes D., Kiss G: Szótárazási dilemmák a MetaMorpho magyar-angol fordítóprogram névszói adatbázisának építésében. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 180–189